Prediction of speech masking release for fluctuating interferers based on the envelope power signal-to-noise ratio

Jørgensen, Søren; Dau, Torsten

Published in:
Proceedings of Acoustics 2012

Publication date:
2012

ACOUSTICS 2012
HONG KONG

Prediction of speech masking release for fluctuating interferers based on the envelope power signal-to-noise ratio

Søren Jørgensen[1] *and Torsten Dau*[2]

Center for Applied Hearing Research, Department of electrical Engineering, Technical University of Denmark,

Ørsteds Plads, Bld. 352, DK-2800 Kgs. Lyngby, Denmark

Abstract: The speech-based envelope power spectrum model (sEPSM) presented by Jørgensen and Dau [(2011). J. Acoust. Soc. Am. **130**, 1475-1487] estimates the envelope signal-to-noise ratio (SNR_{env}) after modulation-frequency selective processing. This approach accurately predicts the speech intelligibility for normal-hearing listeners in conditions with additive stationary noise, reverberation, and nonlinear processing with spectral subtraction. The latter condition represents a case in which the standardized speech intelligibility index and the speech transmission index fail. However, the sEPSM is limited to conditions with stationary interferers due to the long-term estimation of the envelope power and cannot account for the well-known phenomenon of speech masking release. Here, a short-term version of the sEPSM is described [Jørgensen and Dau, 2012, in preparation], which estimates the SNR_{env} in short temporal segments. Predictions obtained with the short-term sEPSM are compared to data from Kjems *et al.* [(2009). J. Acoust. Soc. Am. **126** (3), 1415-1426] where speech is mixed with four different interferers, including speech-shaped noise, bottle noise, car noise, and a highly non-stationary cafe noise. The model accounts well for the differences in intelligibility observed for the stationary and non-stationary interferers, demonstrating further that the SNR_{env} is crucial for speech comprehension.
Keywords: Speech, Intelligibility, Modeling, Masking-release

1. Introduction

The masking of speech is greatly influenced by the spectral and temporal characteristics of the masker. In an early study, Miller and Licklider (1950) demonstrated that the amount of speech masking caused by a stationary noise decreases if the noise is periodically interrupted, while keeping the long-term spectrum and the signal-to-noise ratio (SNR) the same. The corresponding reduction of the speech reception threshold (SRT) for the speech presented in interrupted noise, relative to the threshold in stationary noise, was denoted as speech masking release (MR). The MR was attributed to the listener's ability to take advantage of the temporal dips of the masker with a favorable short-term SNR.

Classical speech intelligibility prediction metrics like the articulation index (AI, ANSI-1969) and, its successor, the speech intelligibility index (SII; ANSI-1997) fail to account for the data, since they are based on the frequency contents of the speech and the noise only. The speech transmission index (STI, IEC-60268) includes information about the temporal modulation energy of the target signal, which typically decreases when stationary noise or reverberation is added. However, the STI does not distinguish between masker and target fluctuations and thus cannot account for the MR phenomenon.

[1] sjor@elektro.dtu.dk
[2] tdau@elektro.dtu.dk

A common characteristic of the classical models is that their predictions are based on the long-term (several seconds) estimates of speech and noise and thus lack fine temporal resolution. To overcome this limitation, Rhebergen and Versfeld (2005) proposed a short-time version of the SII, denoted as the extended SII (ESII). The main element of this metric is the determination of the SII in a number of short temporal windows covering the duration of a given speech token. While this approach was shown to account for various aspects of the MR effect, substantial discrepancies between predictions and data were observed (e.g., when speech was used as the masker). Furthermore, as with the original SII and STI, the ESII fails in conditions with nonlinearly processed noisy speech, such as after amplitude compression (Rhebergen et al., 2009) and noise reduction (Smeds et al., 2011). Thus, the short-term SNR, as calculated in the ESII-metric, does not seem to represent an appropriate indicator of speech intelligibility in adverse conditions.

Recently, it was shown that the change in SRT caused by nonlinear processing of noisy speech could be accounted for using a metric based on the envelope power signal-to-noise ratio (SNR_{env}; Jørgensen and Dau, 2011). The key step in this approach is the determination of the SNR_{env} from the envelopes of noisy speech and noise alone at the output of a modulation-frequency selective process. The SNR_{env} is computed as part of the speech-based envelope power spectrum model (sEPSM; Jørgensen and Dau, 2011), which was shown to account for the intelligibility of different speech materials in stationary noise. In addition, the sEPSM could account for the shift of the SRT caused by applying additional reverberation and processing by spectral subtraction. However, the sEPSM presented in Jørgensen and Dau (2011), is limited to conditions with stationary interferers since the SNR_{env} is computed from the long-term integration of the envelope power. In order to extend the framework to conditions with fluctuating interferers, a short-term version is presented here, where the SNR_{env} is estimated in short temporal windows.

The hypothesis is that the SNR_{env} is a general metric of the intelligibility of speech in both steady and fluctuating noise. Following this hypothesis, it was expected that the SNR_{env} increases during the periods where the masker amplitude is low. This was tested here by comparing model predictions to data obtained using two different speech materials and six different interferers with widely different spectral and temporal characteristics.

2. Model description

The processing structure of the short-term sEPSM is illustrated in Fig. 1A. The first stage is a bandpass filterbank comprised of 22 gammatone filters with ERB bandwidth and one-third octave spacing, covering the range from 63 Hz to 8 kHz. An absolute sensitivity threshold is included such that individual gammatone-filters are included, only if the level of the stimulus at the output is above the absolute hearing threshold for normally hearing listeners. The temporal envelope of each output is extracted via the Hilbert-transform and then low-pass filtered with a cut-off frequency of 150 Hz using a first-order Butterworth filter. The resulting envelope is analyzed by a modulation bandpass filterbank, which consists of eight second-order bandpass filters with octave spacing covering the range from 2 - 256 Hz, in parallel with a third-order lowpass filter with a cutoff frequency of 1-Hz.

The temporal output of each modulation filter is divided into short segments using rectangular windows with no overlap. The durations of the windows are specific for each channel, and set equal to the inverse of the center-frequency of a given bandpass filter (or the cut-off frequency in the case of the 1-Hz low-pass filter). For example, the window duration in the 4-Hz modulation channel is 256 ms. For each segment, the AC-coupled envelope power (variance) of the noisy speech (P_{S+N}) and the noise alone (P_N) are calculated and normalized with the corresponding DC-power. The envelope power of a segment is assumed to have a minimum value of -30 dB (rel. to DC), which reflects the minimum human sensitivity to amplitude modulation (Ewert and Dau, 2000). The SNR_{env} of a segment is estimated from the envelope power as:

$$SNR_{env} = \frac{P_{S+N} - P_N}{P_N} \quad (1)$$

For each modulation channel, the resulting segmental SNR_{env}-values are averaged across all windows covering the duration of a given speech token. The SNR_{env}-values are then combined across modulation filters and across gammatone filters using the "integration model" from Green and Swets (1988). The overall SNR_{env} is converted to the probability of correctly recognizing the speech item using the concept of a statistically "ideal observer". The ideal-observer stage contains two parameters that reflect the response set size and the redundancy of a given speech material.

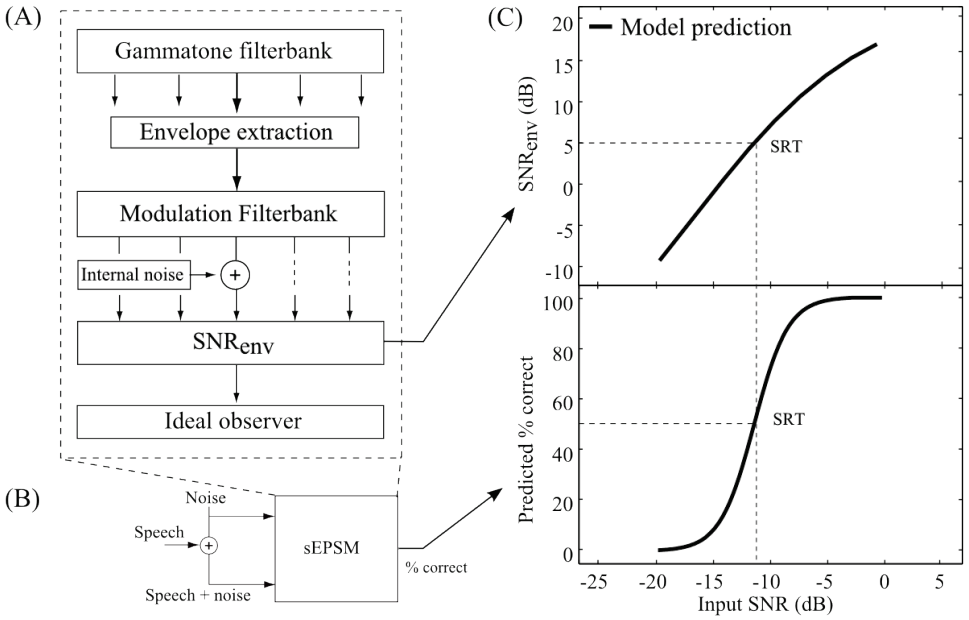

Figure 1 – **(A)** Block-diagram of the sEPSM processing structure. **(B)** Scheme for predicting speech intelligibility using the sEPSM. **(C)** SNR_{env} as a function of the input SNR (top panel) and the corresponding probability of a correct response (bottom panel).

The scheme for predicting intelligibility of noisy speech is shown in Fig. 1B. Noisy speech and noise alone (assumed to be available separately) are used as inputs to the sEPSM. Here, the noise alone represents an estimate of the intrinsic noise within the noisy speech. These inputs are analyzed separately until the SNR_{env}-stage. Figure 1C illustrates the internal representation of the SNR_{env} as a function of the input SNR (top panel) and the corresponding probability of recognizing the speech input (bottom panel). The latter can be regarded as the model's psychometric function. The predicted SRT is defined here as the 50%-point of this psychometric function.

3. Method

Model predictions are compared to two sets of data obtained for normal-hearing (NH) listeners. The first data set is from Kjems *et al*. (2009) and consists of sentences from the Danish version of the Hagerman sentence test, denoted as DANTALE II (Wagener *et al*., 2003). The sentences are mixed with four different interferers: a stationary speech-shaped noise (SSN), a conversation between two people sitting in a Café (Cafe), a car-cabin noise (Car), and the sound of bottles on a conveyer belt (Bottle). The long-term frequency spectra of the maskers are shown in Fig. 2 (Left). While the Café-noise and the SSN have identical long-term spectra, they have different temporal characteristics, since the Café-noise is real speech. Otherwise the frequency spectra of the maskers are very different.

The second data set was obtained using the Danish Conversational Language Understanding Evaluation (CLUE) test (Nielsen and Dau, 2009), which is similar to the HINT-test. SRTs were measured using three maskers: a stationary speech-shaped noise (SSN), a fluctuating noise constructed by fully modulating the SSN with an 8-Hz sinusoid (8-Hz mod), and the non-semantic International Speech Test Signal (ISTS, Holube *et al*., 2010). The SRTs were obtained for five NH listeners using the adaptive

procedure described in Jørgensen and Dau (2011).

Model predictions were obtained as the average across 100 and 150 sentences from the DANTALE II-material and the CLUE-material, respectively. For both speech materials, the model parameters were calibrated to a close match between the predictions and the data for the SSN-conditions. These parameters were then used for all masker-conditions for a given speech material.

4. Results

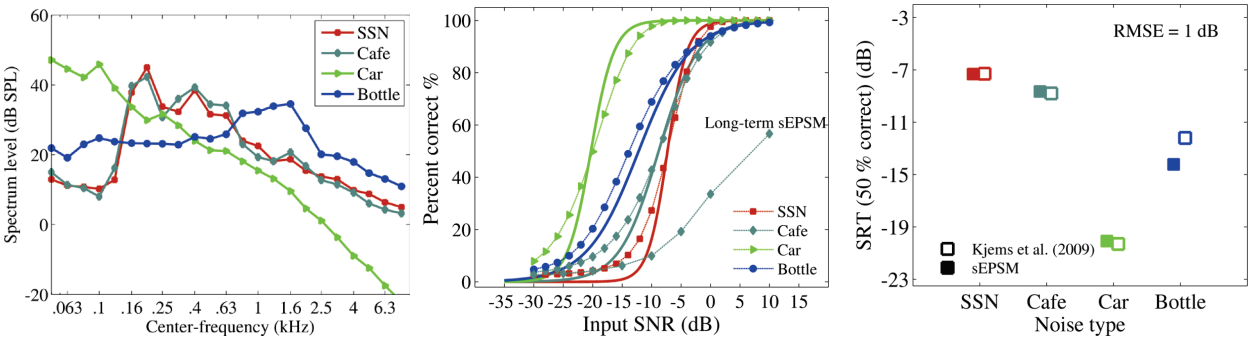

Figure 2 – **Left**: Long-term third-octave frequency spectra of the maskers. **Middle**: psychometric functions (solid lines) estimated from measured data by Kjems *et al.* (2009) and corresponding predictions by the short-term sEPSM (connected symbols) for the four noise conditions. Predictions from the long-term sEPSM with the Café-noise are also shown labeled on the graph. **Right**: Measured and predicted SRTs as a function of the noise type.

The middle panel of Figure 2 shows the psychometric functions (solid lines) measured by Kjems et al. (2009). The corresponding sEPSM predictions are shown as closed symbols connected by straight lines. In addition, predictions from the long-term sEPSM for the Café-noise are shown with a label on the graph. There is a good match between the predicted and the measured for all noise types. However, the predicted slopes are generally slightly shallower. The predictions from the long-term sEPSM clearly fail in the case of the Cafe noise. The model parameters were calibrated only to the SSN-condition. Thus, since the model parameters are fixed, the changes in the predicted intelligibility across the noise-types result only from the differences in the stimuli.

The right panel of Figure 2 shows the measured (open squares) and the predicted (filled squares) SRTs for the four interferers. The lowest SRT for the listeners is obtained with the Car-noise. This can be explained by the frequency content of this noise, which does not mask the speech as effectively as the SSN or the Bottle-noise. The SRT for the Café-noise is slightly lower than for the SSN, even though these two maskers have the same long-term spectrum; this reflects a small MR-effect. A model such as the SII, which only considers spectral information, would not be able to account for this difference. In contrast, the short-term sEPSM accounts well for the SRTs obtained with all four interferers.

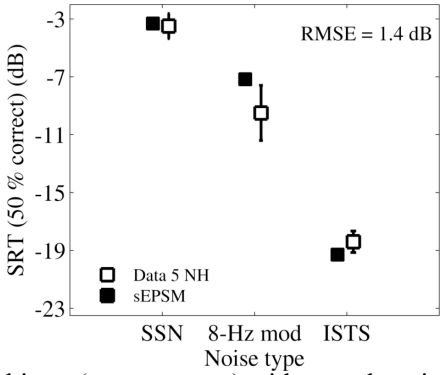

Figure 3 – Average SRTs for five NH subjects (open squares) with error-bars indicating plus/minus one standard deviation and sEPSM predictions (closed squares).

Figure 3 shows the measured (open squares) and predicted (filled squares) SRTs obtained with the CLUE-sentences. In the measured data, the SRTs for the two fluctuating noises were found at lower SNRs than the SSN. The lowest SRT was observed for the speech-like ISTS, consistent with the data by Festen and Plomp (1990). The SRT for the 8-Hz mod-noise is about 6 dB lower than the SRT for the SSN, reflecting a clear MR-effect. Since the long-term spectrum is the same for both maskers, the reduced SRT can be only be explained by the temporal fluctuations of the 8-Hz mod-noise. The lower SRT for the ISTS compared to the SSN is also most likely due to the temporal fluctuations. However, the long-term spectra of these two noise types were not matched, which means that differences in spectral masking may also play a role. The sEPSM-predictions are in agreement with the data, although, the match is not as good as that found with the DANTALE II-material.

5. Model analysis

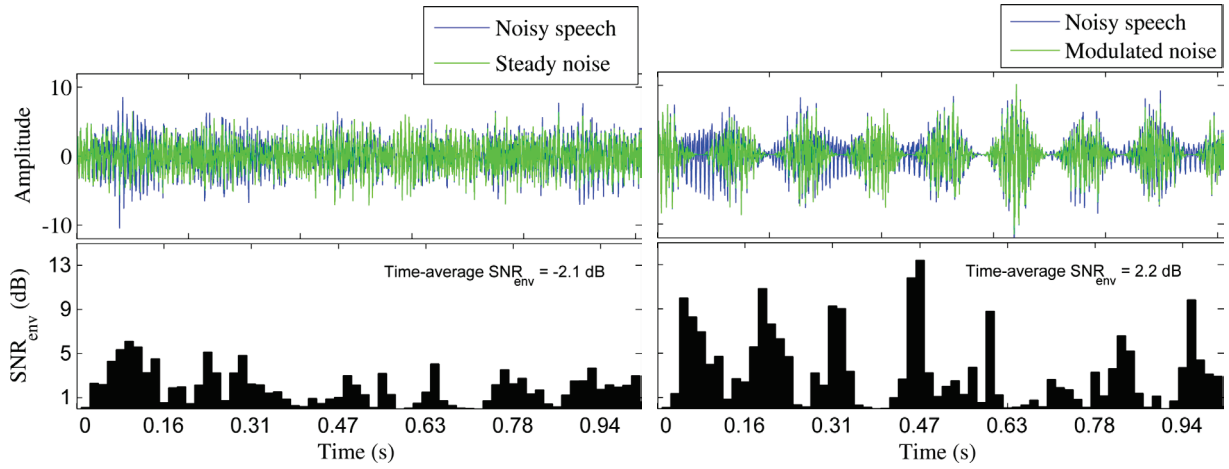

Figure 4 – **Left**: Temporal waveform (top panel) of a sentence mixed with a stationary noise (blue) together with the noise alone (green) and the corresponding SNR_{env} measured in the 64-Hz channel (bottom panel). **Right**: The same situation as the left panel but with speech mixed with a modulated noise.

To illustrate how the prediction of speech MR is reflected in the internal representation of the sEPSM, the top-left panel of Figure 4 shows an example of a sentence-plus-SSN waveform (blue) overlaid by the noise alone (green), both used as inputs to the sEPSM. The top-right panel of Figure 4 shows the same sentence, but here the noise is amplitude modulated at 8 Hz. In both cases, the long-term SNR is -5 dB. The bottom panels of Figure 4 show the corresponding short-term SNR_{env}-values, measured at the output of the 64-Hz modulation filter and averaged across all the audio filters. A comparison of the left and right panels reveals that the SNR_{env} is increased during the dips of the modulated noise and decreased during the peaks. Overall, this leads to an increase of the time-averaged SNR_{env} across the whole sentence, which in turn leads to an increase of the predicted intelligibility. Hence, the MR effect in the model is caused by an increase of the short-term SNR_{env} during the dips of the fluctuating noise.

6. Summary and discussion

Jørgensen and Dau (2011) showed that the sEPSM could account for the change of the SRTs caused by applying reverberation or spectral subtraction to noisy speech. However, the sEPSM, as presented in Jørgensen and Dau (2011), has shortcomings in conditions with fluctuating maskers, since predictions are based on the long-term envelope power. A solution was presented here in the form of a short-term version of the sEPSM. This model accounts well for the SRTs reported by Kjems *et al*. (2009), measured in four noise-conditions with maskers having very different spectral and temporal characteristics. Neither the STI nor the SII perform as well in these tasks (Christiansen *et al*., 2010). Furthermore, the short-term sEPSM predictions were in agreement with the SRTs obtained with the CLUE-sentences in stationary noise, 8-Hz

modulated noise, and for a competitive talker represented by the ISTS material. In contrast, the short-term ESII fails in similar conditions (Rhebergen and Versfeld, 2005).

The ESII and other short-term prediction metrics predict the MR-effect from the short-term SNR in the audio-domain. In these models, the MR-effect results from an increased audio-domain SNR in the dips of the masker. In contrast, the MR predicted by the sEPSM is caused by an increase of the SNR_{env} in the dips of the masker.

Even though other short-term prediction metrics, such as the ESII, can predict some aspects in the MR-data, none of them can account for the effects of nonlinear processing with spectral subtraction. This indicates that the SNR in the audio-domain does not appropriately capture the relevant information in the stimuli underlying the intelligibility of speech in noise. The sEPSM-framework can account for the MR-effect; the key difference between the sEPSM and other intelligibility models is the determination of the SNR in the envelope domain, after frequency selective processing. This supports the hypothesis that the SNR_{env} may capture relevant features underlying the intelligibility of noisy speech.

References and links

Christiansen C., Pedersen, M.S., Dau, T. (2010). "Prediction of speech intelligibility based on an auditory preprocessing model", Speech. Commun., 52, 678–692.

Festen, J. M. and Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing" J. Acoust. Soc. Am. 88 (4), 1725–1736.

Green, D. M. and Swets, J. A. (1988). Signal Detection Theory and Psychophysics (Peninsula Publishing, Los Altos California), 238-239.

Holube, I., Fredelake, S. Vlaming, M., Kollmeier, B. (2010). "Development and analysis of an International Speech Test Signal (ISTS)," Int. J. Audiol. 49(12), 891-903.

Jørgensen, S. and Dau, T. (2011). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing" J. Acoust. Soc. Am., 130 (3), 1475–1487.

Kjems, U., Boldt, J.B., Pedersen, M.S., Lunner, T., Wang, D., (2009). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech", J. Acoust. Soc. Am. 126 (3), 1415–1426.

Miller, G. A., Licklider, J. C. R. (1950). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. 22(2), 167-173.

Nielsen, J. B. and Dau, T. (2009). "Development of a Danish speech intelligibility test", Int. J. Audiol. 48, 729-741.

Rheebergen, K. S., Versfeld,. N. J. (2005). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners", J. Acoust. Soc. Am. 117 (4), 2181–2192.

Rheebergen, K. S., Versfeld,. N. J., Dreschler, W. A. (2009). "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise", J. Acoust. Soc. Am. 126 (6), 3236–3245.

Smeds, K., Wolters, F., Leijon, A., Nilsson, A., Båsjö, S., Hertzman, S. (2011). "Predictive measures of the intelligibility of speech processed by noise reduction algorithms", Proceedings of ISAAR, in press.

Wagener, K., Josvassen, J. L., and Ardenkjaer, R. (2003). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. 42, 10–17.